MycoKeys 26: 13–24 (2017)
doi: 10.3897/mycokeys.26.14591
http://mycokeys.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
MycoKeys
Launched to accelerate biodiversity research

Read quality-based trimming of the distal ends of public fungal DNA sequences is nowhere near satisfactory

R. Henrik Nilsson[1,2], Marisol Sánchez-García[3], Martin Ryberg[4], Kessy Abarenkov[5], Christian Wurzbacher[1,2], Erik Kristiansson[6]

1 *Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden* **2** *Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Göteborg, Sweden* **3** *Department of Biology, Clark University, 950 Main St., Worcester, MA 01610-1477, USA* **4** *Department of Organismal Biology, Uppsala University, Norbyv. 18D, 75236 Uppsala, Sweden* **5** *Natural History Museum, University of Tartu, Vanemuise 46, Tartu 51014, Estonia* **6** *Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden*

Corresponding author: *R. Henrik Nilsson* (henrik.nilsson@bioenv.gu.se)

Academic editor: *I. Schmitt* | Received 19 June 2017 | Accepted 28 July 2017 | Published 14 August 2017

Citation: Nilsson RH, Sánchez-García M, Ryberg M, Abarenkov K, Wurzbacher C, Kristiansson E (2017) Read quality-based trimming of the distal ends of public fungal DNA sequences is nowhere near satisfactory. MycoKeys 26: 13–24. https://doi.org/10.3897/mycokeys.26.14591

Abstract

DNA sequences are increasingly used for taxonomic and functional assessment of environmental communities. In mycology, the nuclear ribosomal internal transcribed spacer (ITS) region is the most commonly chosen marker for such pursuits. Molecular identification is associated with many challenges, one of which is low read quality of the reference sequences used for inference of taxonomic and functional properties of the newly sequenced community (or single taxon). This study investigates whether public fungal ITS sequences are subjected to sufficient trimming in their distal (5' and 3') ends prior to deposition in the public repositories. We examined 86 species (and 10,584 sequences) across the fungal tree of life, and we found that on average 13.1% of the sequences were poorly trimmed in one or both of their 5' and 3' ends. Deposition of poorly trimmed entries was found to continue through 2016. Poorly trimmed reference sequences add noise and mask biological signal in sequence similarity searches and phylogenetic analyses, and we provide a set of recommendations on how to manage the sequence trimming problem.

Key words

Molecular identification, DNA barcoding, database curation, Sanger sequencing, high-throughput sequencing, molecular ecology

Copyright *R. Henrik Nilsson et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Introduction

Molecular (DNA-based) species identification is the process by which newly generated DNA sequences are examined for taxonomic affiliation and sometimes functional aspects by comparison to reference sequences of firmly established taxonomic origin. It is a powerful tool to identify organisms, particularly those with few or no discriminatory morphological characters and those with cryptic or inconspicuous life styles (Pečnikar and Buzan 2014). Fungi are one such group (Stajich et al. 2009). Molecular exploration of substrates such as soil, water, and even household dust from the built environment has revealed a spectacular diversity of hitherto unrecognized fungal lineages (Grossart et al. 2016, Nilsson et al. 2016, Abarenkov et al. 2016), and recent estimates put the number of extant fungal species at upwards of 6 million (Blackwell 2011, Taylor et al. 2014). The number of recognized, validly described species, in contrast, stands at ~135,000 (http://www.speciesfungorum.org, July 2017). Fruiting bodies or other tangible somatic structures are not known for any of the 40 previously unknown fungal lineages examined by Tedersoo et al. (2017), and at present DNA-based methods represent the only way to approach the taxonomic affiliation of these and other lineages.

Several factors combine to make molecular identification of fungi complicated. In addition to the lack of reference sequences for more than 99% of the estimated number of extant species of fungi, technical complications such as chimera formation and low read quality may introduce noise and bias to such efforts (Hyde et al. 2013, Kõljalg et al. 2013). To some extent, software tools are available to exercise some degree of control over these complications (e.g., Edgar et al. 2011, Bengtsson-Palme et al. 2013). Furthermore, many – perhaps even most – researchers seem to be aware of the need to approach existing as well as newly generated sequences in a critical way (e.g., Nilsson et al. 2012, Alm Rosenblad et al. 2016), which nevertheless does not appear to prevent substandard entries from being deposited in the databases of the International Nucleotide Sequence Database Collaboration (INSDC: GenBank, ENA, and DDBJ, Schoch et al. 2014, Cochrane et al. 2016). Such substandard INSDC entries may skew research efforts through, e.g., BLAST sequence similarity searches (Altschul et al. 1997) or inclusion in multiple sequence alignments and phylogenetic analyses.

One aspect of sequence reliability that remains largely unexplored is quality trimming of the distal (approximately 25 bases at the very 5' and 3') ends of Sanger sequences. Owing to the nature of the Sanger sequencing process, the very first bases are often hard to resolve due to the presence of un-incorporated nucleotides and leftover primers. Similarly, the signal-to-noise ratio typically drops with the length of the amplicon in that it becomes increasingly difficult to separate amplicons of near-identical lengths from each other on the electrophoresis gel. Thus, an important part of Sanger sequencing is to inspect the resulting chromatograms and remove any noisy distal sequence parts in the newly generated sequence data. This step is, however, sometimes overlooked. When working with INSDC data for fungal molecular identification and sequence analysis purposes, we regularly come across entries whose distal ends appear to be very poorly trimmed. They may feature extended homopolymer regions (e.g.,

AAAAAAAAA...) or stretches of seemingly random bases that are not found in other conspecific sequences (Nilsson et al. 2012). These potentially noisy sequence ends make it difficult to judge BLAST results: are the mismatches in the distal ends of sequences due to actual biological (nucleotide) differences, or is the reason for the mismatches simply low read quality owing to poor trimming of the reference sequences? There is no direct way of knowing, although clues can perhaps be gleaned from the BLAST alignment and comparison with other conspecific sequences. We fear that, in many cases, researchers will not ponder this question, but will rather assume (or will use automated sequence processing tools that assume) that the mismatches observed are of a biological nature. This will translate into compromised molecular identification, suboptimal assignment of taxonomic affiliations, and unsatisfactory use of sequence data.

The problem is of particular concern for the nuclear ribosomal internal transcribed spacer (ITS) region, the formal fungal barcode and the most popular genetic marker for assessing the taxonomic composition of fungal communities (Schoch et al. 2012, Lindahl et al. 2013). This marker is used by hundreds of studies annually, such that the ramifications of poorly trimmed reference sequences could taint the results of numerous studies each year (cf. Gilks et al. 2002). In an effort to assess the extent of poor sequence trimming in the public sequence repositories, we compared the ITS sequences from 86 fungal (draft) genomes with the public fungal ITS sequences from the same species in the INSDC. We found that in many cases, researchers do not seem to have applied stringent sequence trimming; indeed, in many cases, researchers do not appear to have inspected the chromatograms at all before depositing the sequence data in the INSDC. We conclude by offering a set of observations and recommendations to alleviate the sequence trimming problem in present and future molecular research efforts.

Materials and methods

Retrieval of reference ITS sequences from genomes

The ribosomal operon is regularly left out from genome sequencing efforts due to assembly difficulties (Schoch et al. 2014, Hibbett et al. 2016), such that there is no straightforward way to obtain the ITS region from all existing fungal genomes (as has been reported for other genes, Bai et al. (2015)). We therefore used BLAST in the NCBI Whole Genome Shotgun database (https://www.ncbi.nlm.nih.gov/genbank/wgs/) to identify fully assembled ribosomal regions, using the very conserved 5.8S gene of the ITS region as the BLAST seed. Of the 130 matches returned, 86 were found to represent full-length ITS regions of distinct species that were also represented by at least one reasonably full-length Sanger-derived ITS sequence in the INSDC. In addition to the full ITS region, we kept 50 bases of the upstream nuclear ribosomal small subunit (nSSU/18S) gene and 50 bases of the downstream nuclear ribosomal large subunit (nLSU/28S) gene in the genome-derived ITS sequences to guide the subsequent alignment step.

Retrieval of INSDC sequences

For each of the 86 species (spanning 3 fungal phyla and 29 orders, Suppl. material 1), we downloaded all reasonably full-length ITS sequences from the INSDC using the NCBI query phrase "Species name[ORGN] AND 5.8S[TITL] AND 200:900[SLEN]". We were specifically interested in sequences generated using the traditional ITS1/ITS1F and ITS4/ITS4B primer sets (cf. Tedersoo et al. 2015) since sequences of this coverage are frequently used in DNA barcoding and systematics efforts (Lindahl et al. 2013). Each set of conspecific sequences (the genome-derived sequence plus the conspecific INSDC sequences) was aligned separately in MAFFT 7.307 (Katoh and Standley 2013), and sequences found to contain more than 50 bases of SSU or more than 50 bases of LSU were excluded. Similarly, sequences found to lack more than 50 bases of the 5' end of the ITS1 region, or more than 50 bases of the 3' end of the ITS2 region, were discarded. Alignments were adjusted manually, as needed, following Hyde et al. (2013). Sequences found to be chimeric, taxonomically misidentified, or the subject of other severe technical complications were removed from the alignments prior to statistical analysis. Wherever we found evidence of significant taxonomic variation (e.g., cryptic species) in the alignments, we removed all sequences (alleles) that we deemed to come from a different cryptic species/allele compared to the genome-derived sequence in question. In this study we sought to compare sequence variation in the context of poor sequence trimming rather than in the context of major sequencing artifacts, cryptic species, or allelic divergence.

Multiple sequence alignment and analysis

We went through each position in each of the alignments, starting from the 50th-to-last base of the SSU to the 50th base of LSU, and noted the proportion of INSDC sequences that produced a different nucleotide base from that of the corresponding genome-derived ITS sequence. All three of DNA base mismatches, gaps, and DNA ambiguity symbols (Cornish-Bowden 1985) were counted as mismatches. For each sequence in the alignment, we calculated the dissimilarity (proportion of mismatches) as a function of its relative position. The dissimilarities of the 86 species were then combined using a weighted average with weights proportional to the total number of available sequences for each species. The standard errors were calculated based on the corresponding weighted sample standard deviation. To examine the average age (NCBI date of last modification) of the poorly trimmed sequences, all sequences with at least 5% average dissimilarity among the 5% of its first bases or 5% of its last bases were classified as "potentially poorly trimmed", and their date of NCBI modification was assessed. The association between year and proportion of "potentially poorly trimmed" sequences was examined using overdispersed Poisson rate regression. The relative number of "potentially poorly trimmed" sequences was used as the response variable and time (year) as covariate. All statistical analyses were done in R 3.2.1 (R Core Team 2017).

Results

Multiple sequence alignment

The 86 multiple sequence alignments, each covering at most 50 bases of the SSU, the full ITS region (minus at most 50 bases of the 5' end of ITS1 and/or 50 bases of the end of ITS2), and at most 50 bases of the LSU, are provided in Suppl. material 2. The average length of the alignments was 648 bases (SD: 90, min: 416, max: 941), and the average number of sequences was 123 (SD: 219, min: 1, max: 1586).

Read quality variation

The plotting of disagreements with respect to the genome-derived sequences revealed that insufficient trimming of sequence data seems to be a widespread problem (Figs 1–2). An average of 13.1% of the sequences (SD: 19.6%) in each alignment were classified as "potentially poor trimmed", i.e. they showed at least 5% dissimilarity compared to the corresponding genome sequence over the first or last 5% of the aligned bases. For the remaining (non-distal) bases, those values were down to 0.22% (SD: 0.90%). The dissimilarity was found to be 7.9% and 5.3% in the 5' and 3' ends, respectively (Fig. 2b–c). The proportion of potentially poorly trimmed sequences was consistently high over the years 1997-2016, with a weak but significantly increasing trend ($p=0.0291$, Fig. 3).

Discussion

We provide data to suggest that many public DNA sequences are poorly trimmed in their distal parts. The fact that poorly trimmed sequences continue to be deposited through 2016 furthermore suggests that this problem will not go away by itself over time. We hope that the present paper will serve as an eye-opener, both for researchers who risk using the poorly trimmed data for molecular identification and for researchers generating and depositing sequences in public sequence repositories. The way it is now, these sequences may confound sequence similarity searches by falsely suggesting that two sequences (biological entities) are less similar than what really is the case. This reduces the precision in taxonomic and functional assessment – whether manual or carried out through some software pipeline – of newly generated sequences. Other kinds of sequence analysis, such as phylogenetic analyses, will similarly be distorted by poorly trimmed sequences.

Fortunately, managing read quality in Sanger sequences is fairly straightforward. The chromatograms, indicating the relative signal strength for each of the four purines/pyrimidines C, T, A, and G for each position in the sequence, are a key resource in this pursuit. Brief guidelines for how chromatograms should be processed are available in Hyde et al. (2013) and through various textbooks, online tutorials, and troubleshooting guides (e.g., Kearse et al. 2012, Green and Sambrook 2012). Trying to squeeze out

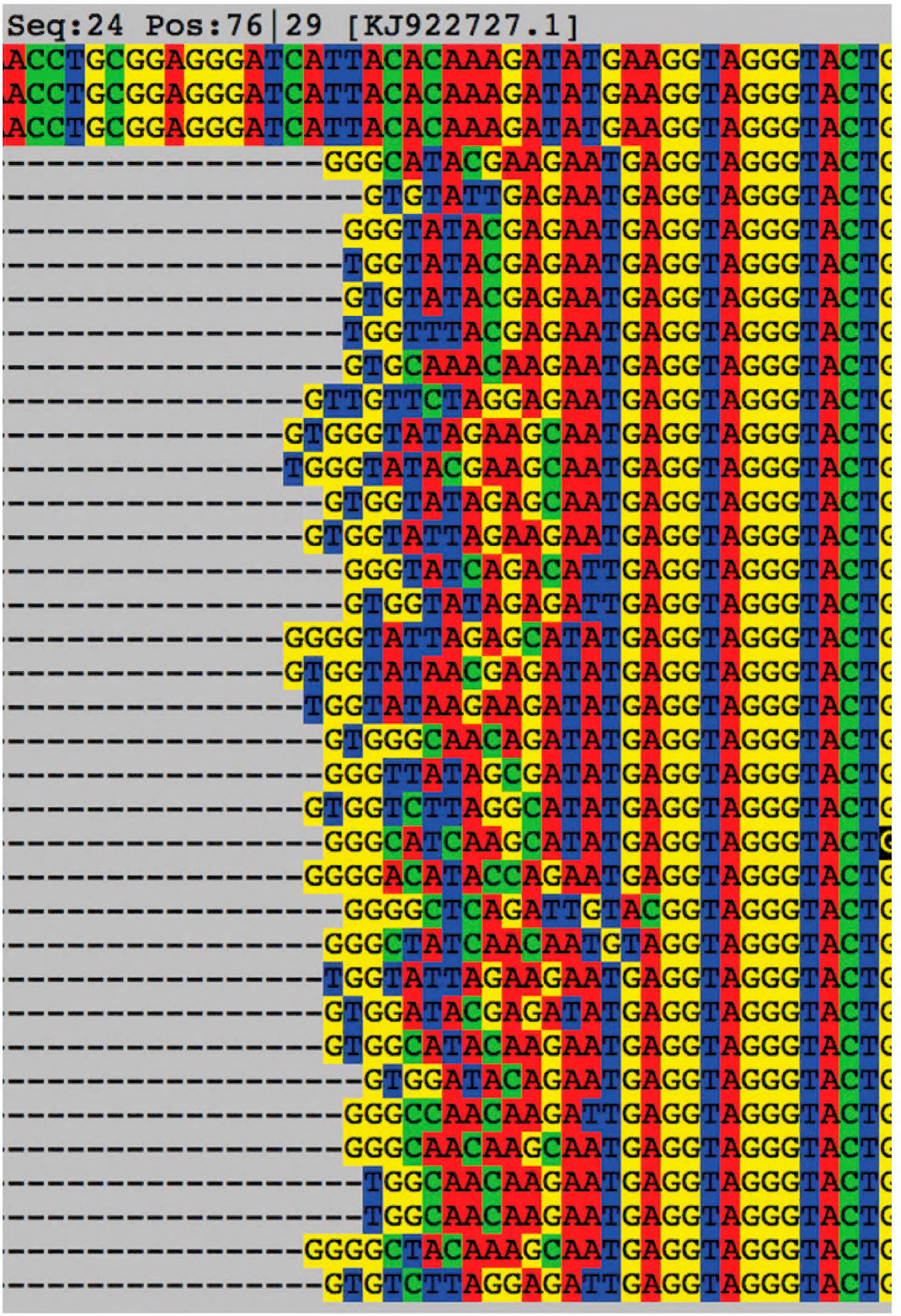

Figure 1. Example of poorly trimmed sequences (sequence four and on) from the species *Setosphaeria turcica*. The 5' end of the alignment is shown, and the poorly trimmed sequences cover the last ~5 bases of SSU and the immediate start of ITS1. The topmost sequence is genome-derived, and sequences two and three are regular Sanger sequences retrieved from the INSDC from other studies than the one with the poorly trimmed sequences (sequences four and on). SeaView v. 4 (Gouy et al. 2010) was used to visualize the alignment.

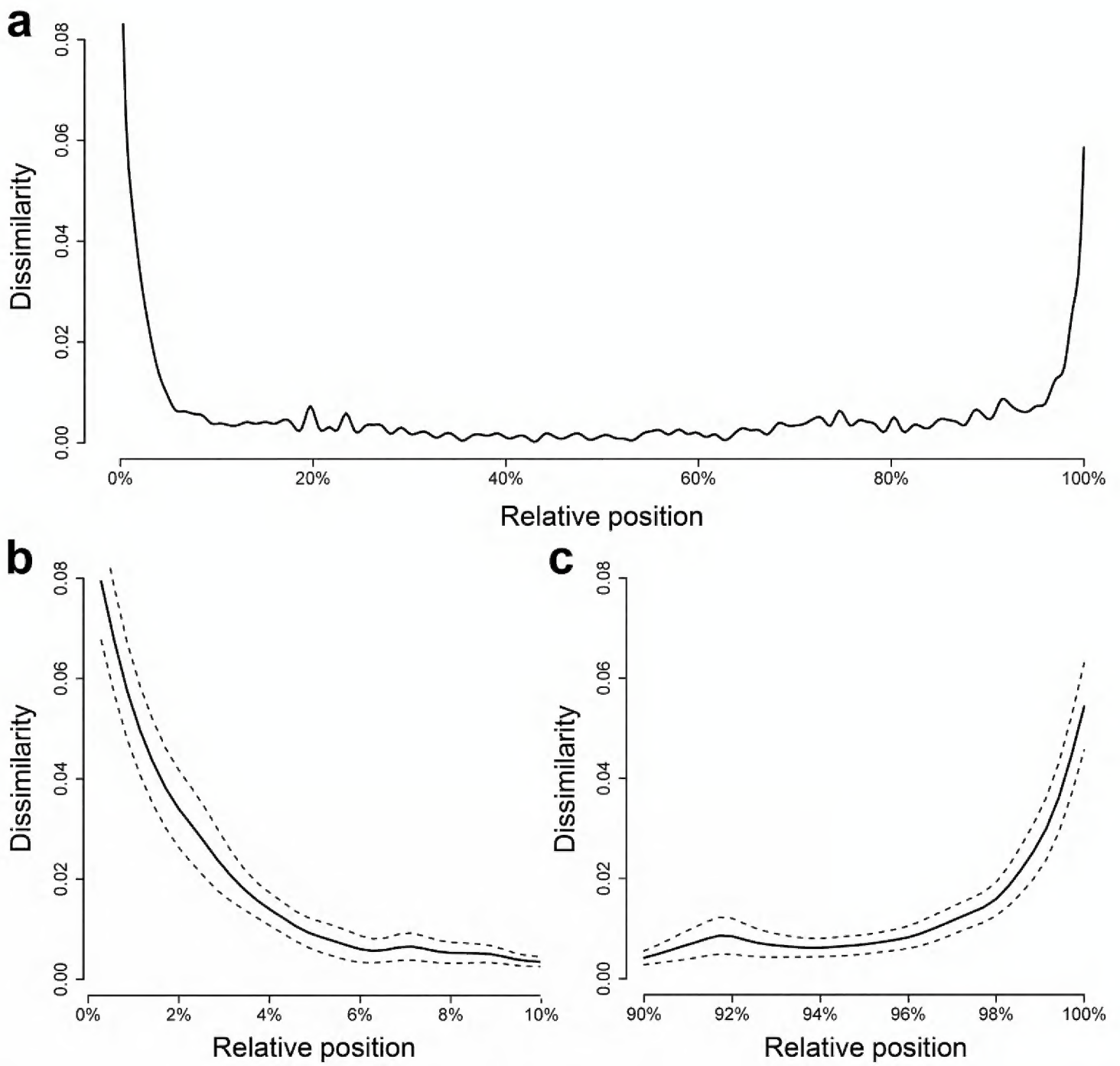

Figure 2. Public fungal ITS sequences are subjected to insufficient trimming in their distal ends. Panel a shows the dissimilarity (y-axis) as a function of the relative sequence position (x-axis). The plot is based on 10,584 sequences from 86 species. Panel b and c show zoom-ins of the 5' and 3' ends, respectively. Dashed lines indicate point-wise standard errors.

extra information from chromatograms by progressing too far in the 5' or 3' ends is not a good idea, and researchers should make it a habit to crop sequence ends aggressively. Generally speaking, habitually trying to salvage sequences with chromatograms of modest overall quality is not likely to be in the best interest of science. In most cases, it would appear to be better to re-process and re-sequence the material using other DNA extraction protocols, primers, or PCR conditions (cf. Larsson and Jacobsson 2004, Young et al. 2014, Lorenz et al. 2017). Finally, sequence similarity searches using BLAST may be used to get an idea of the technical quality of newly generated sequences (cf. Hyde et al. 2013), including at least cursory inspection of whether the distal ends of sequences are trimmed well enough. BLAST is, however, a somewhat blunt tool when it comes to assessing the read quality of sequence ends and we recommend it as a complement to,

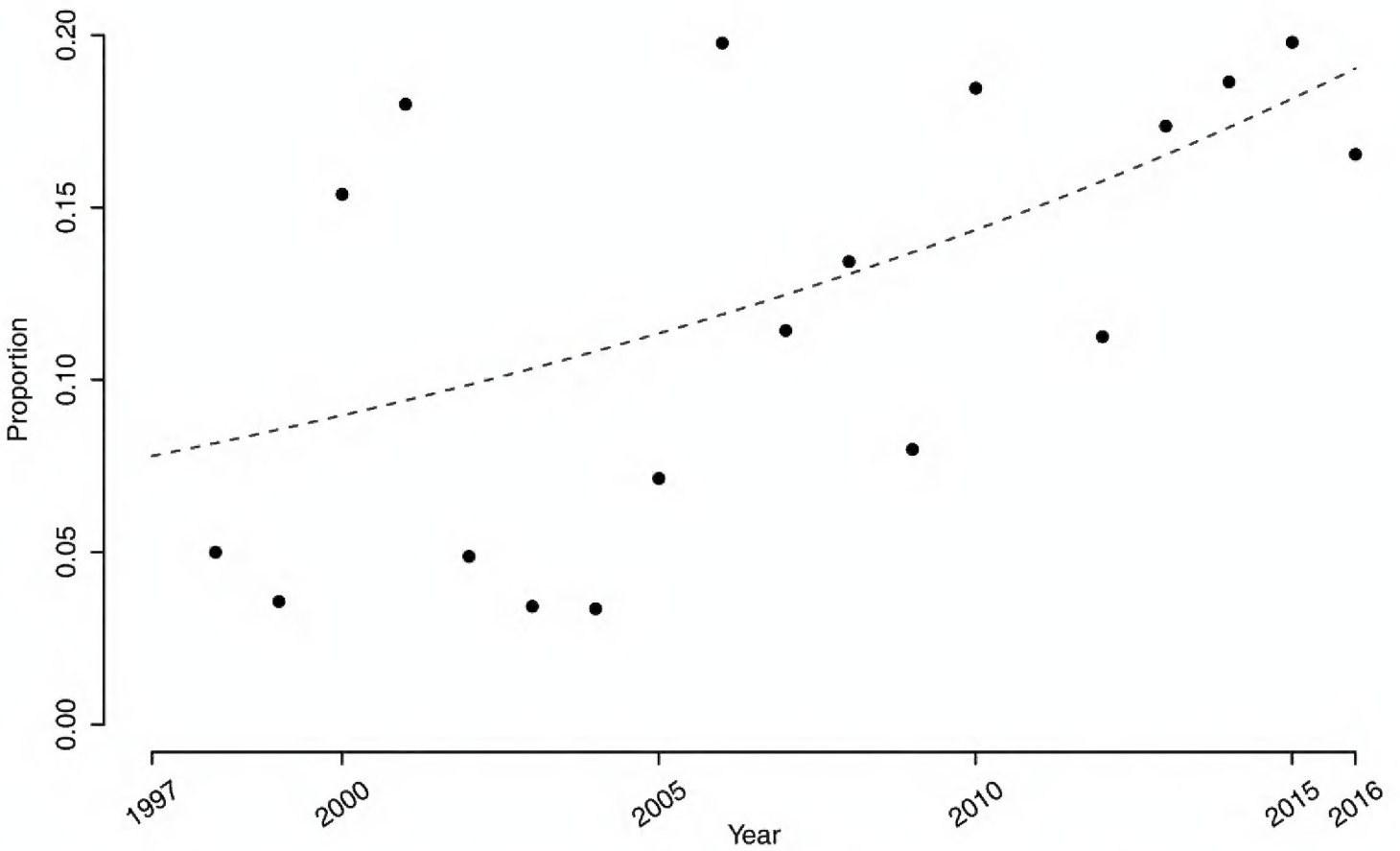

Figure 3. The proportion of poorly trimmed (y-axis) fungal ITS sequences submitted to the INSDC does not decrease over time (x-axis). The regression line (dashed), which was derived by overdispersed Poisson rate regression, shows a weak but significant increasing trend (yearly relative increase of 0.047, p=0.0291).

rather than as a replacement of, manual inspection of chromatograms. NCBI recently launched a unified system for multiple rRNA submission types, the Submission Portal (https://submit.ncbi.nlm.nih.gov/). This includes an ITS submission wizard specifically tailored to provide various verification steps that should decrease the likelihood of low quality submissions. This includes the use of ITSx (Bengtsson-Palme et al. 2013) to improve annotation, vector screening and automatic trimming, plus/minus mis-assembly checks, and trimming or removal of sequences with a high number of ambiguities. Hopefully, this will raise the awareness on part of sequence authors of the need to screen sequence data for quality issues prior to deposition.

In this study we show that incomplete (or lack of) trimming of sequence ends remains abundant in molecular mycology. Although this was expected based on our experience, this is the first study to provide at least an initial estimate of the magnitude of the problem. We used genome-derived ITS sequences from 86 fungal species from 29 different orders in our pursuit, such that we think that it is reasonable to extrapolate our findings to the fungal kingdom at large. Furthermore, we cannot think of any reason why this would be a uniquely fungus-specific problem, and we consider that our findings in fact may hold true for Sanger sequences from all genes and groups of organisms, possibly excluding groups and genes that only a few meticulous researchers have worked on. We would, however, like to stress that we provide estimates rather than hard facts. Our approach relied on genome-derived ITS sequences, and we quantified deviations from the genome sequences among conspecific ITS sequences in the INSDC as assessed through species names (Latin binomials). However, some degree

of deviation from the genome-derived sequences is expected, since intraspecific ITS variation may reach 3% or in some cases more (Schoch et al. 2012, Garnica et al. 2016). Similarly, the multicopy nature of the ITS region is a potential complication in that we may inadvertently have used a rare and perhaps deviant genome ITS copy and compared it to more common ITS copies (cf. Lindner et al. 2013). That said, such intraspecific or intragenomic variation is not known to be limited to the very start and end of the ITS region or other genetic markers and should not be able to produce the pattern seen in Figs 1 and 2a. In addition, we explicitly sought to avoid comparing sequences across different alleles and cryptic species by excluding sequences that did not match the respective genome-derived sequence closely.

In conclusion, we have shown beyond reasonable doubt that there is room for improvement in the way the mycological community – and to some degree the scientific community at large – trim their DNA sequences. The poor sequence trimming leaves a mark on all subsequent studies that make use of those sequences through BLAST searches or otherwise. Mycology faces enough challenges as it is without having to worry about the burden of poorly trimmed sequences (cf. Pautasso 2013), and we hope that this study will serve as a wake-up call when it comes to trimming of sequence entries in mycology and elsewhere.

Acknowledgements

RHN acknowledges financial support from the Swedish Research Council of Environment, Agricultural Sciences, and Spatial Planning (FORMAS, 215-2011-498) and MR from the same agency (FORMAS, 226-2014-1109). RHN, KA, and the UNITE community acknowledge support from the Alfred P. Sloan Foundation. EK acknowledges funding from FORMAS and Wallenberg. CW and RHN acknowledges funding from Stiftelsen Olle Engkvist, Stiftelsen Lars Hiertas Minne, Stiftelsen Birgit och Birger Wåhlströms minnesfond för den bohuslänska havs- och insjömiljön, and Kapten Carl Stenholms donationsfond. Conrad Schoch (NCBI) is gratefully acknowledged for valuable comments on the manuscript. CW acknowledges a Marie Skłodowska-Curie postdoctoral grant (CRYPTRANS).

References

Abarenkov K, Adams RI, Irinyi L et al. (2016) Annotating public fungal ITS sequences from the built environment according to the MIxS-Built Environment standard – a report from a May 23-24, 2016 workshop (Gothenburg, Sweden). MycoKeys 16: 1–15. https://doi.org/10.3897/mycokeys.16.10000

Alm Rosenblad M, Martin MP, Tedersoo L et al. (2016) Detection of signal recognition particle (SRP) RNAs in the nuclear ribosomal internal transcribed spacer 1 (ITS1) of three lineages of ectomycorrhizal fungi (Agaricomycetes, Basidiomycota). MycoKeys 13: 21–33. https://doi.org/10.3897/mycokeys.13.8579

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17): 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Bai L, Xie T, Hu Q et al. (2015) Genome-wide comparison of ferritin family from Archaea, Bacteria, Eukarya, and Viruses: its distribution, characteristic motif, and phylogenetic relationship. The Science of Nature 102(9-10): 64. https://doi.org/10.1007/s00114-015-1314-3

Bengtsson-Palme J, Ryberg M, Hartmann M et al. (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods in Ecology and Evolution 4(10): 914–919. https://doi.org/10.1111/2041-210X.12073

Blackwell M (2011) The Fungi: 1, 2, 3… 5.1 million species? American Journal of Botany 98(3): 426–438. https://doi.org/10.3732/ajb.1000298

Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database Collaboration (2016) The International Nucleotide Sequence Database Collaboration. Nucleic Acids Research 44(D1): D48–D50. https://doi.org/10.1093/nar/gkv1323

Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Research 13(9): 3021–3030. https://doi.org/10.1093/nar/13.9.3021

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27(16): 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

Garnica S, Schön ME, Abarenkov K et al. (2016) Determining threshold values for barcoding fungi: lessons from *Cortinarius* (Basidiomycota), a highly diverse and widespread ectomycorrhizal genus. FEMS Microbiology Ecology 92(4): fiw045. https://doi.org/10.1093/femsec/fiw045

Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA (2002) Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics 18(12): 1641–1649. https://doi.org/10.1093/bioinformatics/18.12.1641

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27(2): 221–224. https://doi.org/10.1093/molbev/msp259

Green MR, Sambrook J (2012) Molecular cloning: a laboratory manual. 4th ed, CSH Press, USA.

Grossart HP, Wurzbacher C, James TY, Kagami M (2016) Discovery of dark matter fungi in aquatic ecosystems demands a reappraisal of the phylogeny and ecology of zoosporic fungi. Fungal Ecology 19: 28–38. https://doi.org/10.1016/j.funeco.2015.06.004

Hibbett D, Abarenkov K, Kõljalg U et al. (2016) Sequence-based classification and identification of Fungi. Mycologia 108(6): 1049–1068. https://doi.org/10.3852/16-130

Hyde KD, Udayanga D, Manamgoda DS et al. (2013) Incorporating molecular data in fungal systematics: a guide for aspiring researchers. Current Research in Environmental and Applied Mycology 3(1): 1–32. https://doi.org/10.5943/cream/3/1/1

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30(4): 772–780. https://doi.org/10.1093/molbev/mst010

Kearse M, Moir R, Wilson A et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12): 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kõljalg U, Nilsson RH, Abarenkov K et al. (2013) Towards a unified paradigm for sequence-based identification of Fungi. Molecular Ecology 22(21): 5271–5277. https://doi.org/10.1111/mec.12481

Larsson E, Jacobsson S (2004) Controversy over *Hygrophorus cossus* settled using ITS sequence data from 200 year-old type material. Mycolological Research 108(7): 781–786. https://doi.org/10.1017/S0953756204000310

Lindahl BD, Nilsson RH, Tedersoo L et al. (2013) Fungal community analysis by high-throughput sequencing of amplified markers - a user's guide. New Phytologist 199(1): 288–299. https://doi.org/10.1111/nph.12243

Lindner DL, Carlsen T, Nilsson RH, Davey M, Schumacher T, Kauserud H (2013) Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer (ITS) rDNA region in fungi. Ecology and Evolution 3(6): 1751–1764. https://doi.org/10.1002/ece3.586

Lorenz MG, Lustig M, Linow M (2017) Fungal-grade reagents and materials for molecular analysis. Methods in Molecular Biology 1508: 141–150. https://doi.org/10.1007/978-1-4939-6515-1_6

Nilsson RH, Tedersoo L, Abarenkov K et al. (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4: 37–63. https://doi.org/10.3897/mycokeys.4.3606

Nilsson RH, Wurzbacher C, Bahram M et al. (2016) Top 50 most wanted fungi. MycoKeys 12: 29–40. https://doi.org/10.3897/mycokeys.12.7553

Pautasso M (2013) Fungal under-representation is (indeed) diminishing in the life sciences. Fungal Ecology 6(5): 460–463. https://doi.org/10.1016/j.funeco.2013.03.001

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences USA 85(8): 2444–2448.

Pečnikar ŽF, Buzan EV (2014) 20 years since the introduction of DNA barcoding: from theory to application. Journal of Applied Genetics 55(1): 43–52. https://doi.org/10.1007/s13353-013-0180-y

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Schoch CL, Robbertse B, Robert V et al. (2014) Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. Database (Oxford) vol. and 10.1093/database/bau061. https://doi.org/10.1093/database/bau061

Schoch CL, Seifert KA, Huhndorf S et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences USA 109(16): 6241–6246. https://doi.org/10.1073/pnas.1117018109

Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW (2009) The fungi. Current Biology 19(18): R840–855. https://doi.org/10.1016/j.cub.2009.07.004

Taylor DL, Hollingsworth TN, McFarland JW, Lennon NJ, Nusbaum C, Ruess RW (2014) A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. Ecological Monographs 84(1): 3–20. https://doi.org/10.1890/12-1693.1

Tedersoo L, Anslan S, Bahram M et al. (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10:1–43. https://doi.org/10.3897/mycokeys.10.4852

Tedersoo L, Bahram M, Puusepp R, Nilsson RH, James TY (2017) Novel soil-inhabiting clades fill gaps in the fungal tree of life. Microbiome 5(1): 42. https://doi.org/10.1186/s40168-017-0259-5

Young JM, Rawlence NJ, Weyrich LS, Cooper A (2014) Limitations and recommendations for successful DNA extraction from forensic soil samples: a review. Science and Justice 54(3): 238–244. https://doi.org/10.1016/j.scijus.2014.02.006

Supplementary material 1

Details on the fungal genomes/contigs targeted

Authors: R. Henrik Nilsson, Marisol Sánchez-García, Martin Ryberg, Kessy Abarenkov, Christian Wurzbacher, Erik Kristiansson

Data type: Excel spreadsheet

Explanation note: List of the fungal genomes/contigs targeted, their URL, their taxonomic affiliation, and the number of sequences (with and without poor trimming) for each entry.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mycokeys.26.14591.suppl1

Supplementary material 2

The 86 multiple sequence alignments used

Authors: R. Henrik Nilsson, Marisol Sánchez-García, Martin Ryberg, Kessy Abarenkov, Christian Wurzbacher, Erik Kristiansson

Data type: Text

Explanation note: The multiple sequence alignments used to infer the statistics of the study. They are provided in the FASTA format (Pearson and Lipman 1988). The genome-derived sequence is given as the topmost sequence in each alignment.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mycokeys.26.14591.suppl2